CLAIMS

1. A method for identifying a set of genes from a multiplicity of genes whose expression levels at a first state and a second state are measured in replicates using one or more nucleotide arrays, thereby generating a first plurality of independent measurements of the expression levels for said first state and a second plurality of independent measurements of the expression levels for said second state, which method comprises:

(a) identifying a quality function capable of evaluating the distinctiveness between the first plurality and the second plurality;

(b) selecting a subset of genes, whose expression levels in said first state and second state are represented in said first plurality and said second plurality, respectively;

(c) calculating the values of the quality function for said subset of genes in said first state and said second state based on the first and second plurality, thereby determining the distinctiveness of the first and the second plurality;

(d) substituting a gene in said subset with one outside of said subset, thereby generating a new subset, and repeating step (c), keeping the new subset if the distinctiveness increases and the original subset if otherwise;

(e) repeating steps (c) and (d) for a first predetermined number of times, thereby identifying a locally optimal subset of genes;

(f) repeating steps (b) to (e) for a second predetermined number of times, thereby identifying said second predetermined number of the locally optimal subsets; and

(g) integrating said second predetermined number of the locally optimal subsets into said set of genes, wherein said set is larger than said locally optimal subsets in size.

2. The method of claim 1, wherein said states are selected from the group consisting of biological states, physiological states, pathological states, and prognostic states.

3. A method for identifying a set of genes from a multiplicity of genes whose expression levels at a first tissue and a second tissue are measured in replicates using one or more nucleotide arrays, thereby generating a first plurality of independent measurements of the expression levels for said first tissue and a second plurality of independent measurements of the expression levels for said second tissue, which method comprises:

(a) identifying a quality function capable of evaluating the distinctiveness between the first plurality and the second plurality;

(b) selecting a subset of genes, whose expression levels in said first tissue and second tissue are represented in said first plurality and said second plurality, respectively;

(c) calculating the values of the quality function for said subset of genes in said first tissue and second tissue based on the first and second plurality, thereby determining the distinctiveness of the first and the second plurality;

(d) substituting a gene in said subset with one outside of said subset, thereby generating a new subset, and repeating step (c), keeping the new subset if the distinctiveness increases and the original subset if otherwise;

(e) repeating steps (c) and (d) for a first predetermined number of times, thereby identifying a locally optimal subset of genes;

(f) repeating steps (b) to (e) for a second predetermined number of times, thereby identifying said second predetermined number of the locally optimal subsets; and

(g) integrating said second predetermined number of the locally optimal subsets into said set of genes, wherein said set is larger than said locally optimal subsets in size.

4. The method of claim 3, wherein said tissues are selected from the group consisting of normal lung tissues, cancer lung tissues, normal heart tissues, pathological heart tissues, normal and abnormal colon tissues, normal and abnormal renal tissues, normal and abnormal prostate tissues, and normal and abnormal breast tissues.

5. A method for identifying a set of genes from a multiplicity of genes whose expression levels in a first type of cells and a second type of cells are measured in replicates using one or more nucleotide arrays, thereby generating a first plurality of independent measurements of the expression levels for said first type of cells and a second plurality of independent measurements of the expression levels for said second type of cells, which method comprises:

(a) identifying a quality function capable of evaluating the distinctiveness between the first plurality and the second plurality;

(b) selecting a subset of genes, whose expression levels in said first type of cells and said second type of cells are represented in said first plurality and said second plurality, respectively;

(c) calculating the values of the quality function for said subset of genes in said first type of cells and said second type of cells based on the first and second plurality, thereby determining the distinctiveness of the first and the second plurality;

(d) substituting a gene in said subset with one outside of said subset, thereby generating a new subset, and repeating step (c), keeping the new subset if the distinctiveness increases and the original subset if otherwise;

(e) repeating steps (c) and (d) for a first predetermined number of times, thereby identifying a locally optimal subset of genes;

(f) repeating steps (b) to (e) for a second predetermined number of times, thereby identifying said second predetermined number of the locally optimal subsets; and

(g) integrating said second predetermined number of the locally optimal subsets into said set of genes, wherein said set is larger than said locally optimal subsets in size.

6. The method of claim 5, wherein said types of cells are selected from the group consisting of normal lung cells, cancer lung cells, normal heart cells, pathological heart cells, normal and abnormal colon cells, normal and abnormal renal cells, normal and abnormal prostate cells, and normal and abnormal breast cells.

7. The method of claim 5, wherein said type of cells are selected from the group consisting of cultured cells and cells isolated from an organism.

8. The method of claim 1, 3, or 5, wherein said integrating is performed by selecting the genes whose frequency of occurrences in said second predetermined number of the final subsets exceeds a third predetermined number.

9. The method of claim 8, wherein said third predetermined number is 1% or 5%.

10. The method of claim 1, 3, or 5, wherein said first predetermined number is sufficiently small such that the global maximum is not reached.

11. The method of claim 1, 3, or 5, wherein said quality function is a parametric function or a non-parametric function.

12. The method of claim 11, wherein said parametric function is selected from the group consisting of the Mahalanobis distance and the Bhattacharya distance.

13. The method of claim 1, 3, or 5, wherein the nucleotide arrays are selected from the group consisting of arrays having spotted thereon cDNA sequences and arrays having synthesized thereon oligonucleotides.